**Customizations of Clustering Algorithms for Mixed Data:**

**Need, Challenges and Prospects**

 *Pradnya Bhambre1 Dr. Mrs. Nusrat Khan2*

pradnyabhambre@gmail.com nusrat.khan@sinhgad.edu

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**I Abstract**

In real life, there are many datasets which involve mixed types of data i.e. categorical data and numerical data. These datasets can be from the healthcare domain, finance domain, marketing domain, [11] business analytics or any other prominent field. To enhance the quality of the outcome and to get new insights in these datasets one has to search trends, correlations and dependencies between data variables etc. This can be achieved by analyzing the data by using some appropriate and effective method. But analysis of such data is very difficult, since any one available clustering algorithm could not be directly applied to these datasets for data analysis with desired efficiency. Many attempts have been made to apply some clustering algorithm to this mixed data. For this, some customizations or modifications based on different criteria have been applied to the traditional clustering algorithms. These customizations result in an algorithm or method which could analyze mixed data more effectively, may provide better accuracy and may improve time complexity.

1. Research Scholar, Savitribai Phule Pune University

2. Associate Professor, Sinhgad Institutes, Pune

This research paper explores the need for these customizations, illustrates challenges faced in applying these customizations to the traditional clustering algorithms and enlists possible prospects of mixed data clustering in various data domains.

**Keywords** clustering algorithms, mixed data, categorical data, healthcare data

**II Introduction**

Any real-world dataset has mixed type data. Some variables in this dataset are numeric as well as some variables are categorical. To interpret results from this data, there is a need to investigate the relationships between different variables. Some clustering algorithms can be used effectively to cluster pure numeric data and some can be used to cluster pure categorical data. But the performance of these traditional clustering algorithms is poor for mixed categorical and numeric data. [1,7] Therefore, there is a need to modify these clustering algorithms according to the nature of the dataset. But this task is not easy. There are many challenges in this modification process. This paper finds out the need for these customizations, challenges faced in this process and also enlists the prospects or benefits of this customizations.

**III Literature Study**

Shih M., Jheng W. and Lai L., in the research paper “A Two-Step Method for Clustering Mixed Categorical and Numeric Data”, [1] proposed a new two-step clustering method which is useful for finding clusters on mixed type data. It involves investigating the relationships among categorical variables and then converting them into numeric variables. Then the existing distance - based clustering algorithms can be applied to form the clusters. The number of subsets is fixed to one-third of the total number of objects. This method overcomes the weaknesses of k-means clustering algorithm as well as produces high quality clusters. This method integrates k-means clustering and hierarchical clustering algorithms to cluster mixed types of data. Also, applying other clustering algorithms or similarity measures into this method may provide better results.

Sen W., Hong C. and Xiaodong F., “Clustering Algorithm for Incomplete Data Sets with Mixed Numeric and Categorical Attributes”, [2] in this research paper authors have proposed an improved algorithm which overcomes the limitation of traditional k-prototypes algorithm of requirement of data without missing values. This algorithm implements a new dissimilarity measure and a new approach for selecting the initial prototypes based on the nearest neighbors. It finds clusters on the data set and then imputes missing values. The accuracy of this proposed algorithm was compared with traditional k-prototypes algorithm and k-prototypes with the new dissimilarity measure by using data set from the UCI machine learning repository database. This proposed k-prototypes algorithm is superior and has higher accuracy, providing more reasonable clustering results. Also, it minimizes error by better imputing the missing values.

Rauf A., Sheeba, Mahfooz S., Khusro S. and Javed H., “Enhanced K-Mean Clustering Algorithm to Reduce Number of Iterations and Time Complexity”, [3] in this research paper authors proposed a new method of clustering using K-means algorithm which calculates initial centroids instead of randomly selecting them. To overcome the drawback of Traditional K-mean clustering algorithm which requires selection of initial centroids randomly. Therefore, the proposed algorithm has achieved the advantage of reducing the number of iterations and improving the elapsed time.

Ahmad A. and Khan S., “A Novel Initial Clusters Generation Method for K-means-based Clustering Algorithms for Mixed Datasets”, [4] in this paper authors proposed a new approach to find initial partition for K-means-based clustering algorithms for mixed datasets. This k-means-based clustering algorithm is used multiple times and in each run one of the attributes is used to produce the initial partition. Then the clustering results of various iterations are combined to produce an initial partition. This initial partition is considered as initial input for a k-means-based clustering algorithm for clustering mixed data. The initial partitions produced here are always fixed, do not change for different runs or for change in the order of the data objects. Application of this proposed InitKmix algorithm to various mixed datasets shows that it produces accurate and consistent results.

Nguyen D., Nguyen G. and Lam V., “An Approach to Data Mining in Healthcare: Improved K-means Algorithm”, [5] the central idea used in this research paper is ‘Patients having the same health status can be clustered to discover new disease’. But it has a disadvantage that selecting a suitable number of clusters is not easy and obvious. This paper reviews existing methods and presents an improved algorithm for obtaining the number of initial centroids and also provides a new measure for selecting the number of clusters.

Priya Anand R., Jebathangam J. and Bhuvana R., “Approaches to Partition Thyroid Data using Clustering Algorithms”, [6] this research paper provides the implementation of various clustering techniques for thyroid data to ease the design of clustering methods for further enhancement.

Reddy M. and Kavitha B., “Clustering the Mixed Numerical and Categorical Data set using Similarity Weight and Filter Method”, [7] in this paper authors present a clustering algorithm based on similarity weight and filter method for mixed dataset. They proposed a modified description of cluster center to overcome the numeric data only limitation and provide a better characterization of clusters. The performance of algorithm has been studied using some benchmark data sets. And it is concluded that this method is better than the K-Prototype algorithm. This method reduces time complexity and is efficient for any number of dimensions. Also, irregular boundaries can be handled efficiently using Divide and Conquer Technique.

JI J., Pang W., Zheng Y., Wang Z. and Ma Z., “An Initialization Method for Clustering Mixed Numeric and Categorical Data Based on the Density and Distance”, [8] in this research paper, a new initialization method for mixed data clustering is proposed, in which both the distance and density are considered together to determine initial cluster centers. The performance of this method is compared with that of traditional initialization methods using various experiments on three real-world datasets viz. Heart Disease, Zoo and Credit Approval datasets from UCI machine learning repository. But, the disadvantage of this method is, the time complexity is quadratic because of the expensive evaluation of the density for data objects.

Anwar T., Siswantining T., Sarwinda D., Soemartojo S. and Bustamam A., “A study on missing values imputation using K-Harmonic means algorithm: Mixed datasets”, [9] this paper implements K-Harmonic means clustering to reduce random centroid initialization sensitivity problems. Imputation of the missing values is carried out by distributing missing values observation to the cluster and then replacing the missing values with the information on the same centroid cluster. The results of the simulation and the accuracy values of each imputation value for numerical and categorical data were evaluated using the root mean square error.

Khan S. and Ahmad A., “A Cluster center initialization algorithm for K-modes clustering”, [10] in this research paper the author proposed a new cluster center initialization algorithm which ensures fixed initial cluster centers and which provides repeatable clustering results. This method performs better in terms of accuracy and time complexity. The initial cluster centers obtained by this method are close to the actual cluster centers, which provides the faster convergence of K-modes clustering algorithm and better clustering results. The algorithm is composed of two parts 1) Relevant attributes selection and 2) Computing initial cluster centers. The biggest advantage of the proposed method is the worst-case log-linear time complexity of computation and fixed choice of initial cluster centers from dense localized regions. If the number of desired clusters is not available in advance, the proposed multi-clustering approach is extended for the categorical data for finding out the natural number of clusters present in the data, in addition to obtaining the initial cluster centers. It is concluded that considering fewer most relevant attributes is a better choice than selecting all attributes.

Khan S. and Ahmad A., “Survey of State-of-the-Art Mixed Data Clustering Algorithms”, [11] in this survey paper authors identified five major research themes for the study of mixed data clustering and presented a comprehensive literature review. Also, strengths and weaknesses of these methods, the challenges incurred and future directions within each research theme are discussed. And many high-impact areas of applications, possible open research questions, and future research directions are enlisted. This can help researchers to get an in-depth knowledge of the field of clustering mixed data and facilitate the generation of new ideas to solve the real-world problems.

Andreopoulos B., An A. and Wang X., “Bi-level clustering of mixed categorical and numerical biomedical data”, [12] here the authors proposed an algorithm which partitions biomedical data sets of mixed types, such as hepatitis, thyroid disease and yeast gene expression data with Gene Ontology annotations more accurately. This algorithm uses categorical clustering as a prior to maximize the probabilities that objects will be assigned to the correct clusters. This technique useful in the field of evidence-based medicine, for drawing conclusions about the outcome of a patient’s condition based on evidence from outcomes of other patients’ conditions for Physicians. Biologists will also find this method useful in wet lab work, for obtaining hints about the potential functions of genes and proteins.

Lam D., Wei M. and Wunsch D., “Clustering Data of Mixed Categorical and Numerical Type with Unsupervised Feature Learning”, [13] in this paper, unsupervised feature learning is applied to the mixed-type data to achieve a sparse representation, which makes it easier for clustering algorithms to separate the data. It works with the mixed-type data using fuzzy adaptive resonance theory to obtain a better clustering result by removing the differences in treating categorical and numeric features. The advantages of this are demonstrated with several real- world data sets including heart disease, teaching assistant evaluation, and credit approval. This approach is explained using noisy, mixed-type petroleum industry data. UFLA can learn its features even when the amount of data is small in important subspaces of the dataset. Visual assessment tendency is used to determine the true number of clusters in the dataset when the number of clusters is unknown. Results from the application of this method to several real datasets demonstrate the effectiveness of the approach.

Ahmad A. and Dey L., “A k-mean clustering algorithm for mixed numeric and categorical data”, [14] this paper presents a modified k-mean clustering algorithm which uses a new cost function and distance measure for categorical attribute values to compute the distance of an object from a cluster center. These measures consider the significance of an attribute towards the clustering process. It presents a modified description of cluster center to overcome the limitation of numeric data and to provide a better characterization of clusters. The performance of this algorithm is better as compared with other clustering algorithms on real world datasets. It also proposed a modified representation for the cluster center. This representation can effectively capture cluster characteristics, as it contains the distribution of all categorical values in a cluster.

**IV Customizations of Clustering Algorithms:**

**Need:**

In real-world applications, data objects with both numeric and categorical features are ubiquitous. [8] Clustering is often applied to mixed datasets to find structures and to group similar objects for further analysis. [11] The coexistence of both categorical and numerical attributes makes the clustering methods designed for single-type data inapplicable to mixed-type data. [8] Hence, the traditional clustering algorithms are not useful for clustering mixed data. As the data variables in mixed type data are both numeric and categorical, it is difficult to apply same mathematical or statistical operations to find clusters on this data. Either both type of data should be operated differently or firstly the categorical data should be converted into numeric and then the complete dataset can be clustered. Still this conversion may not be exact and thus it may generate some errors. Therefore, researchers need to find some modifications or customizations which can be applied to traditional clustering algorithms, so that these algorithms can be best suited for clustering mixed data. And optimal clustering results can be obtained.

**Challenges:**

Mixed-type categorical and numerical data are a challenge in many applications. [13] To obtain clustering results with desired accuracy in mixed data is difficult. Customizations of the traditional clustering algorithms to get the required results in mixed dataset, faces the following challenges:

1. It is difficult to directly apply mathematical operations, such as summation or averaging, to the categorical values of mixed datasets. [11] So modification of clustering algorithms needs some rule which suits both numeric and categorical variables operations.

2. In the process of Data cleaning, often missing values are found in the dataset. And removing the data containing missing values can reduce information in the dataset. Therefore, this data could not be removed. These missing values can be filled by using imputation with mean, median, mode, regression, likelihood etc. [9] So before clustering mixed dataset with missing values, there is a need to apply some formula or rule to fill these missing values in data preprocessing. Applying this customization is not obvious.

3. Traditional k-means and k-modes clustering algorithms have the drawback of randomly selecting the initial cluster centers [4,8,10] for each run which may lead to non-repeatable clustering results. [4,10] These clustering algorithms can be applied to partition the categorical data into predefined k- clusters, but the clustering results intrinsically depend on the choice of random initial cluster centers, that can cause non-repeatable results and may produce improper cluster structures. Therefore, there is an inconsistency of clustering results. And the inferences obtained from the datasets may be unreliable. [4] Implementing the customizations, to search the initial cluster centers such that it can provide repeatable and consistent results of clustering is a difficult task.

4. There are some initialization algorithms developed to compute initial partition for mixed datasets, but they are computationally expensive or do not produce consistent clustering results in different iterations. [4] Therefore to use these available algorithms with required customizations is neither cost effective nor easy.

5. Some improved clustering algorithms which are formed by applying some customizations to traditional clustering algorithms, may optimize results, overcome the weakness, makes the result more easily analyzable, but may reduce the execution speed. [5] Therefore, when applying customization, the care should be taken that execution speed should not decrease below desired.

6. The success of data clustering often depends on good data, rather than good algorithms. If the dataset is huge and not clear, the choice of clustering algorithm could not provide satisfactory performance, so clustering algorithm based on speed or ease of use, have to be selected. [5] So the required customization should be applied to make data clean and complete. It is difficult for huge datasets.

7. Different modifications in traditional clustering algorithms for clustering mixed data may provide clustering results with different accuracies. Therefore, by applying another modification rule, it is always possible to achieve better performance than before.

**Prospects:**

Biomedical datasets often have mixed categorical and numerical variables, which can be clustered to find out required results. Physicians will find clustering mixed data useful in the field of evidence-based medicine, for drawing conclusions about the outcome of a patient’s condition. Biologists will also find this mixed data clustering useful in wet lab work, such as obtaining clues about the potential functions of genes and proteins. [12]

In the healthcare domain, mixed data clustering can be applied to discover hidden patterns. In Healthcare Management, data mining is useful in decision making [5,6] in treatments, detecting disease diagnosis patterns, fault detection of medical devices, medical fraud prevention and detection, healthcare quality improvement strategies etc. [6]

Mixed data sets occur frequently in many domains, such as health informatics, finance, business analytics, social studies and marketing etc. [11]

In this paper, only the application areas for which researchers already tried to cluster mixed data are enlisted, but these applications are not limited and there can be numerous future prospects for mixed data clustering. Various prominent data fields such as Education, Geology, Economics, Environmental sciences, Biology, Social science, Management, Human resource, Population studies etc. have tremendous uses of clustering mixed type of data. Therefore, new researchers are encouraged to search these applications areas, to apply required customizations to the traditional clustering algorithms and analyze the data to get the required information from the desired data domain. This can provide the amazing results by discovering new perceptions or views about the correlations, trends or dependencies between variables in the data domain under consideration.

**V Conclusion**

In everyday life, we come across various high impact data fields containing mixed data. This paper enlists different customizations that can be applied to traditional clustering algorithms, so that these algorithms can be used to obtain optimal clustering results for the mixed data. These customizations depend on the factor like nature of the dataset, the clustering accuracy needed and the required execution speed, feasibility conditions etc. This paper explains the need of these customizations in detail, challenges incurred in applying these customizations to the traditional clustering algorithms and also it enlists some real-world prospects of these customizations. By understanding all these aspects, some useful customizations can be applied according to the need, and better clustering results can be obtained from mixed data domains. This paper will encourage new researchers to effectively use mixed data clustering in different data domains to get desired clustering analysis.

**VI References**

1. Shih M., Jheng W., Lai, L., “A Two-Step Method for Clustering Mixed Categorical and Numeric Data”, Tamkang Journal of Science and Engineering, Vol. 13, No. 1, pp. 11-19 (2010)
2. Sen W., Hong C., Xiaodong F., “Clustering Algorithm for Incomplete Data Sets with Mixed Numeric and Categorical Attributes”, International Journal of Database Theory and Application Vol.6, No.5 (2013), pp.95-104, http://dx.doi.org/10.14257/ijdta.2013.6.5.09
3. Rauf A., Sheeba, Mahfooz S., Khusro S. and Javed H., “Enhanced K-Mean Clustering Algorithm to Reduce Number of Iterations and Time Complexity”, Middle-East Journal of Scientific Research 12 (7): 959-963, 2012 ISSN 1990-9233,

DOI: 10.5829/idosi.mejsr.2012.12.7.1845

1. Ahmad A. and Khan S., “A Novel Initial Clusters Generation Method for K-means-based Clustering Algorithms for Mixed Datasets”, January 2019 <https://www.researchgate.net/publication/330752154>
2. Nguyen D., Nguyen G., Lam V., “An Approach to Data Mining in Healthcare: Improved K-means Algorithm”, Journal of Industrial and Intelligent Information, Vol. 1, No. 1, March 2013 doi:10.12720/jiii.1.1.14-18
3. Priya Anand R., Jebathangam J. and Bhuvana R., “Approaches to Partition Thyroid Data using Clustering Algorithms”, International Journal of Control Theory and Applications, ISSN: 0974-5572, International Science Press, Volume 10, Number 25, 2017 https://www.researchgate.net/publication/334942566
4. Reddy M. and Kavitha B., “Clustering the Mixed Numerical and Categorical Data set using Similarity Weight and Filter Method”, International Journal of Database Theory and Application, Vol. 5, No. 1, March 2012 https://www.researchgate.net/publication/281445891
5. Ji J., Pang W., Zheng Y., Wang Z., Ma Z., “An Initialization Method for Clustering Mixed Numeric and Categorical Data Based on the Density and Distance”, International Journal of Pattern Recognition and Artificial Intelligence, Vol. 29, No. 7 (2015) 1550024 (16 pages) DOI: 10.1142/S021800141550024X
6. Anwar T., Siswantining T., Sarwinda D., Soemartojo S., Bustamam A., “A study on missing values imputation using K-Harmonic means algorithm: Mixed datasets”, Conference Paper in AIP Conference Proceedings, December 2019, DOI: 10.1063/1.5141651
7. Khan S., Ahmad A., “A Cluster center initialization algorithm for K-modes clustering”, Article in Expert Systems with Applications, December 2013, DOI: 10.1016/j.eswa.2013.07.002
8. Khan S, Ahmad A, “Survey of State-of-the-Art Mixed Data Clustering Algorithms”, IEEE Access, VOLUME 4, 2016, DOI: 10.1109/ACCESS.2017
9. Andreopoulos B., An A., Wang X., “Bi-level clustering of mixed categorical and numerical biomedical data”, Int. J. Data Mining and Bioinformatics, Vol. 1, No. 1, 2006
10. Lam D., Wei M., Wunsch D., “Clustering Data of Mixed Categorical and Numerical Type with Unsupervised Feature Learning”, IEEE Access, VOLUME 3, 2015, DOI:10.1109/ACCESS.2015.2477216
11. Ahmad A., Dey L., “A k-mean clustering algorithm for mixed numeric and categorical data”, Data & Knowledge Engineering 63, 2007, pp. 503–527 Available online at [www.sciencedirect.com](http://www.sciencedirect.com) DOI: 10.1016/j.datak.2007.03.016